UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Estudo de marcadores do cromossoma Y com interesse forense em imigrantes de países africanos no Centro e Sul de Portugal

Ângela Cristina Silva Dente

Mestrado em Biologia Molecular e Genética

Dissertação orientada por:
Prof. Doutor Manuel do Carmo Gomes

2019

Resumo

Portugal continental, sobretudo nas regiões litorais e nomeadamente na região de Lisboa e Vale do Tejo, tem sido uma região atrativa para muitas comunidades de imigrantes, particularmente de origem africana, descendentes ou imigrantes provenientes de antigas colónias Portuguesas em África. Em 2018 residiam em Portugal 477.472 cidadãos estrangeiros, representando cerca 4,6% do total de residentes, sendo estes de várias nacionalidades diferentes, e podendo-se destacar a população oriunda de países africanos (18,8%).

Devido à tendência para o aumento da imigração, a introdução de novos grupos populacionais em Portugal, com características genéticas diferentes, poderá introduzir variabilidade genética na população de acolhimento, que por sua vez pode alterar a estrutura genética da população, nomeadamente, frequências relativas de determinados alelos. Estes fluxos migratórios e a variabilidade genética introduzida nas populações podem interferir com a interpretação de resultados com interesse forense, visto que estes dependem de frequências alélicas conhecidas. O seu estudo é imperativo para garantir que a avaliação da perícia forense seja o mais rigorosa possível e para que estas alterações não tenham impactos significativos em processos de investigação presentes e futuros. Este estudo deve ser efetuado através de marcadores genéticos do cromossoma Y e autossómicos.

O kit PowerPlex® Y23 System (Promega) contém *primers* para amplificação simultânea de 23 marcadores genéticos do cromossoma Y, em particular: DYS576, DYS389I/II, DYS448, DYS19, DYS391, DYS481, DYS549, DYS533, DYS438 (penta), DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643 (penta), DYS393, DYS458, DYS385a/b, DYS456 e Y-GATA-H4. A partir da utilização deste kit para a análise de marcadores genéticos, pode-se obter uma caracterização representativa das populações masculinas imigrantes residentes no sul de Portugal.

Este trabalho tem como objetivo a caracterização genética de uma amostra representativa de indivíduos imigrantes de países africanos residentes na zona sul de Portugal, com marcadores genéticos do cromossoma Y, e respetiva comparação com a população de acolhimento, avaliando o possível impacto da sua introdução na mesma.

Conclui-se que diferenças significativas entre a população de acolhimento e as populações em estudo neste trabalho, logo as suas frequências alélicas e estruturas genéticas deverão também ser diferentes. Este estudo fornece evidência de que a inserção destes haplótipos na população poderá ter um impacto significativo na estrutura genética da população residente no Sul e Centro de Portugal. Tal impacto poderá afetar estudos populacionais ou processos criminais dependentes destas frequências.

Palavras-chave: Genética Forense; Imigração; Cromossoma Y; STRs; PowerPlex® Y23 System

Abstract

Continental Portugal, especially in coastal regions and specifically in the Lisboa e Vale do Tejo region, has been an attractive hub for many immigrant communities, in particular those with African ascendance, descendants or immigrants from African countries that used to be Portuguese colonies. In 2018, Portugal housed 477.472 foreign citizens, representing about 4,6% of the overall resident population, being that these belong to a wide breadth of nationalities, of which the African population stands out considerably (18,8%).

Given the tendency for immigration rates to increase, the introduction of new populational groups in Portugal, with different genetic characteristics, may introduce genetic variability in the host population, which can in turn alter the genetic structure of the population, namely, the relative frequency of certain alleles. These migration fluxes and the genetic variability introduced in populations may affect the interpretation of results in forensic investigations, as these depend on known allele frequencies. The study of these populations is essential to guarantee the utmost accuracy of the evaluation of forensic testing, and that these alterations do not significantly affect ongoing or future forensic investigations. This study may be conducted through the analysis of Y chromosome genetic markers.

The PowerPlex® Y23 System (Promega) amplification kit contains primers for the co-amplification of 23 different Y chromosome genetic markers, in particular: DYS576, DYS389I/II, DYS448, DYS19, DYS391, DYS481, DYS549, DYS533, DYS438 (penta), DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643 (penta), DYS393, DYS458, DYS385a/b, DYS456 e Y-GATA-H4. With the use of this kit, a representative characterisation of the male immigrant populations residing in central and southern Portugal can be obtained.

This study intends to conduct the genetic characterisation of a representative sample of African immigrants residing in central and southern Portugal by means of Y chromosome genetic markers. This sample will then be compared with the host population, in order to evaluate its potential impact in the genetic makeup of the latter.

Here, we conclude that there are differences between the host population and the populations being studied in this work, therefore, their allelic frequencies and genetic structures should also be different. This study offers evidence that the insertion of these haplotypes in the population could then have a significant impact in the genetic structure of the resident population in South and Central Portugal. Such an impact could affect population studies or criminal investigations that depend on these frequencies.

Keywords: Forensic Genetics; Immigration; Y chromosome; STRs; PowerPlex® Y23 System

Agradecimentos

A realização deste mestrado e respetiva dissertação não teria sido possível sem o apoio de todas as pessoas que me acompanharam ao longo desta fase. Gostaria então de agradecer a todos aqueles que contribuíram para a realização dos meus objetivos neste mestrado.

À Doutora Cláudia Vieira Silva, por me ter guiado ao longo da realização de todo o trabalho laboratorial e escrito envolvido neste projeto, bem como na realização de trabalhos adjacentes à dissertação cujas publicações e apresentações contribuíram para o avanço da minha carreira científica.

À Doutora Heloísa Afonso Costa, pelo seu apoio técnico e intelectual ao longo da realização do trabalho laboratorial no Instituto Nacional de Medicina Legal e Ciências Forenses – Delegação do Sul.

À Dona Conceição, pela sua constante atitude positiva e companhia ao longo dos dias passados no Instituto.

Ao meu orientador, Professor Doutor Manuel Carmo Gomes, por ter aceite a posição para esta dissertação.

Por fim, aos meus amigos e família, por me apoiarem ao longo de toda a minha vida académica, mas especialmente durante a realização desta dissertação, e por me darem a coragem e força que necessitei para concluir a mesma.

Os meus mais sinceros agradecimentos a todos.

Índice

Tabelas e Figuras

Tabelas

Figuras

Acrónimos e Símbolos

Single Nucleotide Polymorphisms (SNPs); *Short Tandem Repeats* (STRs); *Y-chromosome Single Nucleotide Polymorphisms* (Y-SNPs); *Y-chromosome Short Tandem Repeats* (Y-STRs); *Y-chromosome Haplotype Reference Database* (YHRD); Índex de fixação (F_{ST}); *Molecular Evolutionary Genetics Analysis* (MEGA); *Multidimensional Scaling* (MDS); *National Institute of Standards and Technology's STRBase database* (NIST).

1. Introdução

1.1. Caracterização de fontes de variabilidade genética no genoma humano

A maior parte do genoma humano é idêntico em todos os indivíduos da espécie. Ainda assim, existem zonas do genoma com elevada variabilidade genética entre indivíduos. Fontes de variabilidade no genoma podem ser mutações ou recombinação genética das quais originam variações em determinadas sequências de DNA encontradas na população, as quais são denominadas de polimorfismos (Hedrick, 2011). Polimorfismos podem ser classificados de diversas formas, de acordo com o tipo de características apresentadas.

Os *Single Nucleotide Polymorphisms* (SNPs) são polimorfismos que correspondem a variações de uma única base numa sequência de DNA, e são uma das maiores fontes de variabilidade no genoma humano. DNA satélite corresponde a outro tipo de polimorfismos, e é constituído por unidades de repetição concatenadas, podendo pertencer a duas categorias: minissatélites e microssatélites. Estas variam no tamanho e no número das unidades de repetição.

Os Microssatélites, também denominados por *Short Tandem Repeats* (STRs), são sequências de DNA repetitivo constituídas por unidades de repetição (de 2-6 pares de base) que se repetem entre 2-50 vezes e representam cerca de 3% do genoma humano, encontrando-se aproximadamente a cada 30-60kbp (Tautz e Renz, 1984; Weber e May, 1989). Nos STRs, os polimorfismos são caracterizados pelo número de repetições do motivo de DNA. Devido a uma taxa de mutação mais elevada associada aos STRs em relação aos SNPs (Brinkmann *et al.*, 1998), estes são uma fonte de variabilidade genética bastante útil em vários contextos. SNPs são frequentemente utilizados em estudos populacionais e estudos de evolução molecular, enquanto STRs são mais utilizados identificação de indivíduos no âmbito médico-legal e forense (Meyer *et al.*, 1995).

Em genética forense são utilizados alguns marcadores genéticos para a identificação de indivíduos em processos variados do âmbito médico-legal e forense, ou mesmo para a investigação de parentesco. Em geral, os marcadores mais utilizados atualmente são STRs, sendo que a sua variabilidade permite um poder de discriminação elevado com uma combinação de apenas 13 marcadores genéticos, enquanto são necessários cerca de 50 SNPs para obter um poder de discriminação semelhante (Glover *et al.*, 2010). Os polimorfismos acima descritos podem ser ou não individualizantes, bem como ser úteis para estudos populacionais a partir da identificação de diferenças características de determinadas populações.

A característica individualizante dos marcadores STR verifica-se apenas em marcadores genéticos autossómicos, visto que a combinação de alelos presentes no genoma de cada indivíduo é uma combinação única dos dois progenitores (com exceção de gémeos idênticos). Os cromossomas sexuais, X e Y, no entanto, são haploides em indivíduos do sexo masculino. Sendo o cromossoma Y um cromossoma haploide, a maior parte do DNA não tem um homólogo com o qual possa haver recombinação genética durante a divisão celular. Devido a este facto, o cromossoma Y é herdado na linha paterna virtualmente sem alterações com exceção de mutações pontuais.

Para além disso o cromossoma Y é o terceiro cromossoma mais curto do genoma humano, sendo apenas ligeiramente mais longo do que os cromossomas 21 e 22. O seu tamanho reduzido contribui para o facto do cromossoma Y estar menos sujeito a degradação, preservando a sua integridade em amostras com sinais de degradação avançada. Isto torna o cromossoma Y útil para análises forenses nas quais as amostras se revelam demasiado degradadas para análise de marcadores genéticos autossómicos após lise diferencial para a obtenção de DNA de células espermatozoides (Roewer, 2009).

Apesar de não serem individualizantes, os marcadores do cromossoma Y são ainda assim bastante úteis para análise de linhas paternas e populacionais devido à sua conservação genética de geração para geração. Estas características tornam o cromossoma Y um excelente 'relógio molecular' para o genoma humano, desde que a sua taxa de mutação seja conhecida (Balanovsky, 2017). A análise de marcadores do cromossoma Y é também particularmente importante para a identificação de misturas de DNA (de mais do que 2 indivíduos masculinos não relacionados pela mesma linhagem paterna), especialmente em casos em que a amostra foi recolhida de um indivíduo do sexo feminino. Uma vez que a presença de marcadores do cromossoma Y não seria esperada numa amostra deste tipo, com a exceção de eventos genéticos raros (eg: XXY), a sua identificação revela a presença de DNA masculino.

Haplótipos são *clusters* de genes herdados em conjunto de um único progenitor. Todo o cromossoma Y, sendo herdado de pai para filho, é considerado um haplótipo. Haplogrupos são grupos de haplótipos semelhantes que podem ser ligados a um ancestral comum através de estudos de genética evolutiva. Certas populações têm então haplogrupos específicos para o cromossoma Y, dependendo da sua origem e das mutações encontradas nos seus haplótipos, o que permite a sua caracterização através da análise de marcadores genéticos presentes neste cromossoma. Devido a estas características, idades estimadas de determinados haplogrupos podem ser ligadas a grandes eventos de migrações populacionais (Balanovsky, 2017).

As mutações são o principal fator responsável pela variabilidade genética de populações. O futuro de uma mutação genética no seio de uma população, depende maioritariamente de três forças evolucionais: *drift* genético aleatório, fluxos migratórios e seleção natural. O primeiro, tal como as mutações, ocorre aleatoriamente e, por isso, as suas consequências para o destino de uma mutação não são previsíveis de forma determinística. No caso da seleção natural, esta atua diretamente a nível do fenótipo, mas tem uma influência indireta no genótipo, visto que a manifestação fenotípica depende do genótipo do indivíduo. Diferentes ambientes podem influenciar de forma diferente a genética das populações que neles habitam através da seleção natural (Cavalli-Sforza, Menozzi e Piazza, 1994).

Os fluxos migratórios são também frequentemente responsáveis por alterações nas frequências alélicas. Fenómenos como a colonização de territórios previamente desabitados, envolvendo apenas uma fração do *pool* genético da população de origem, e a consequente evolução das populações separadas (colonizadores e colonos) resulta em futuras diferenças na estrutura genética das mesmas (Cavalli-Sforza, Menozzi and Piazza, 1994). Da mesma forma, grandes fluxos migratórios em locais com grandes densidades populacionais podem também afetar a estrutura da população de acolhimento a partir de, por exemplo, introdução de alelos anteriormente inexistentes ou que eram considerados raros nesta população. Quer isto dizer que a estrutura genética da população num todo é alterada pela entrada ou saída de *pools* genéticos, o que pode afetar estudos populacionais ou processos criminais dependentes destas frequências.

Para investigar esta possibilidade, é necessário o estudo das populações envolvidas, de modo a compreender a evolução da população de acolhimento após a introdução dos imigrantes aí residentes.

1.2. Caracterização das populações migrantes em Portugal

Em 2018, foi registado um total de população residente em Portugal de 10.276.617 pessoas, das quais 64,5% pertencem à faixa etária de 15-64 anos de idade. Nesse mesmo ano, foram também registados 477.472 imigrantes residentes em Portugal, representando cerca de 4,6% da população total residente no país. A maioria dos imigrantes residentes em Portugal vêm de países Europeus,

representando 42,4% dos mesmos, seguindo-se 24,7% de imigrantes provenientes do continente americano, 18,8% com origem africana e 14% originários da Ásia. Particularmente, destacam-se os imigrantes de países africanos originários maioritariamente de antigas colónias portuguesas, designadamente: Cabo Verde (38,3%), Angola (20,4%), Guiné-Bissau (17,8%), São Tomé e Príncipe (10,5%) e Moçambique (3,3%) (PORDATA, 2018).

É importante investigar as características da população imigrante de origem africana com o objetivo de a comparar com a população do Centro e Sul de Portugal, a fim de avaliar a sua influência na estrutura genética da mesma.

Na Figura 1.1, podemos observar a distribuição de haplogrupos de DNA do cromossoma Y dominantes nas populações nativas de determinadas áreas geográficas. De facto, o continente africano apresenta uma distribuição de haplogrupos de DNA do cromossoma Y bastante diferente dos encontrados na Europa (A, B e E em contraste com I e R, respetivamente) com a exceção de algumas zonas de sobreposição (R1b). É de notar que esta área de sobreposição geográfica dos haplótipos observados não corresponde aos países africanos dos quais se verifica uma maior afluência de imigrantes para Portugal. Isto significa que os haplogrupos tipicamente encontrados em Portugal e nos países de onde são oriundos imigrantes africanos diferem, e fluxos migratórios irão alterar as frequências relativas de certos marcadores genéticos em Portugal.

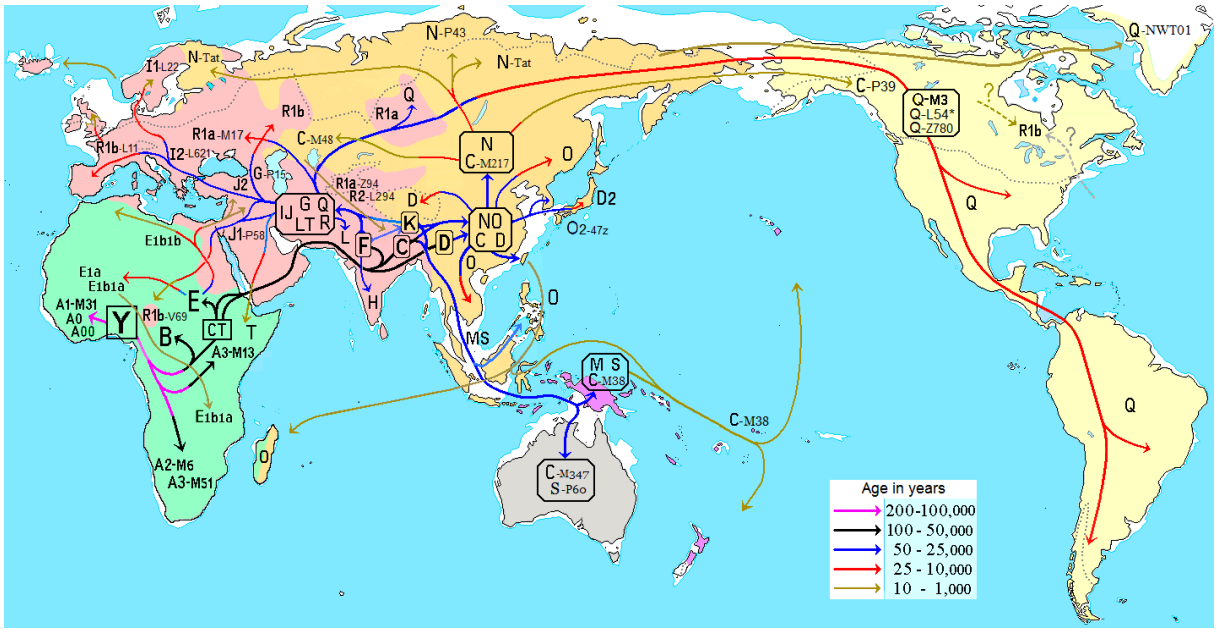

Figura 1.1-Mapa Mundial com possíveis rotas migratórias pre-históricas com base em haplogrupos de DNA do cromossoma Y. Fonte: https://en.wikipedia.org/wiki/File:Migraciones_humanas_en_haplogrupos_de_ADN-Y.PNG. Acedido: Novembro de 2019

Em genética forense, exames de identificação de indivíduos ou de investigação de parentesco, são realizados através da análise de marcadores genéticos e comparação entre perfis genéticos. Os resultados obtidos consistem em probabilidades de perfis genéticos obtidos estarem relacionados (dependendo do tipo de processo é formulada uma hipótese H_0), comparativamente à possibilidade dos mesmos perfis genéticos estarem relacionados com um ou outros indivíduos ao acaso na população (hipótese alternativa H_1), dependendo do tipo de processo em estudo. Para esta avaliação probabilística são utilizadas frequências populacionais dos alelos que caracterizam cada um dos grupos populacionais, pelo que é fundamental o seu estudo. Os fluxos migratórios podem alterar estas frequências, e daí alterar as probabilidades associadas aos exames genéticos realizados no âmbito forense.

1.3. Marcadores genéticos do cromossoma Y

Existem kits de amplificação no mercado adequados para a caracterização de marcadores genéticos do cromossoma Y como Y Filer™ e Y Filer™ Plus (Applied Biosystems) e o PowerPlex® Y e PowerPlex® Y23 (Promega) de entres muitos outros. Estes kits contêm *primers* que permitem a co-amplificação de determinados marcadores genéticos para posterior análise e caracterização. Em particular, o kit de amplificação PowerPlex® Y23 (Promega), permite a co-amplificação de 23 marcadores genéticos do cromossoma Y: DYS576, DYS389I/II, DYS448, DYS19, DYS391, DYS481, DYS549, DYS533, DYS438 (penta), DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643 (penta), DYS393, DYS458, DYS385a/b, DYS456 e Y-GATA-H4. Estes serão os marcadores genéticos caracterizados neste estudo (Figura 1.2).

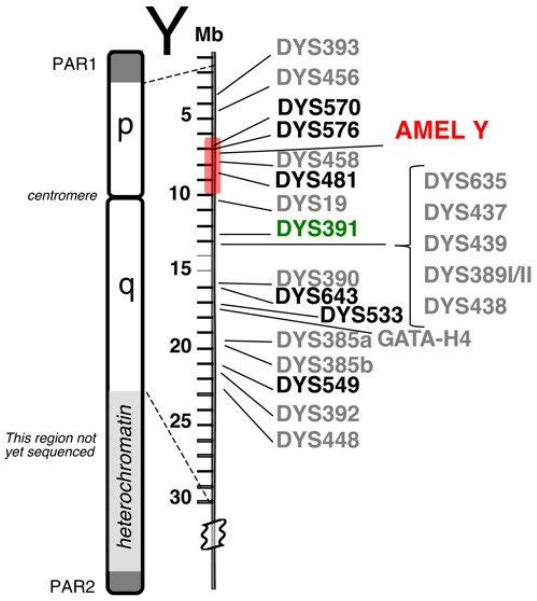

Figura 1.2–Posições relativas dos 23 locus Y-STR amplificados pelo kit PowerPlex® Y23 System. Fonte: https://worldwide.promega.com/resources/profiles-in-dna/2012/variability-of-new-str-loci-and-kits-in-us-population-groups/. Acedido: Novembro de 2018.

Os marcadores DYS389I/II e DYS385a/b são marcadores multi-cópia, querendo isto dizer que há duas cópias destes marcadores STR em zonas diferentes do cromossoma. DYS389II refere-se ao comprimento total do marcador DYS389, sendo DYS389I uma porção mais pequena do marcador em questão. Isto significa que mutações em DYS389I refletem-se no marcador DYS389II. No entanto, DYS385a e DYS385b são cópias independentes e indivíduos com o mesmo perfil de DYS385 podem estar separados por dois ou mais eventos mutacionais.

A análise de marcadores genéticos do cromossoma Y é útil para a identificação de DNA masculino numa mistura em que existe um excesso de DNA feminino em relação a DNA masculino, como é o caso de processos de natureza sexual. Nestas situações, pode-se realizar uma lise diferencial para separar o DNA de espermatozoides do DNA de células epiteliais vaginais, de modo a obter um perfil autossómico do DNA masculino (Gill, Jeffreys and Werrett, 1985). Extrações diferenciais tendem a ser ineficazes quando as amostras em estudo são mais antigas ou parcialmente degradadas, tendo também a desvantagem de eliminar DNA de células masculinas que não sejam espermatozoides.

A análise de Y-STRs amplifica apenas marcadores do cromossoma Y, logo mesmo extraindo todo o DNA presente numa amostra, incluindo DNA feminino, apenas o DNA masculino será detetado (Roewer, 2009). No entanto, determinados eventos cromossómicos durante a replicação como

duplicações, deleções e *slippage* da polimerase de DNA podem também alterar a estrutura de alguns destes marcadores genéticos. Sendo um cromossoma haploide, com exceção de marcadores genéticos identificados como multi-cópia, é de esperar que exista apenas um alelo por marcador.

Duplicações de segmentos do cromossoma poderão resultar na presença de dois alelos em marcadores genéticos caracterizados por apenas um. Em casos de investigação forense, este tipo de duplicações poderá lançar dificuldades na análise de tais amostras, podendo sugerir presença de uma mistura de DNA quando tal não se verifica. A possível presença mais pronunciada de certas duplicações em marcadores genéticos do cromossoma Y em determinadas populações torna a caracterização destas mesmas importante em contextos forenses.

1.4. *Y-chromosome Haplotype Reference Database* – A base de dados de variantes de sequências do cromossoma Y

A *Y-chromosome Haplotype Reference Database* (YHRD) é uma base de dados pública que aloja sequências do cromossoma Y, analisadas em amostras populacionais devidamente documentadas. Esta base tem três objetivos principais: 1) a geração de estimativas de frequência de haplótipos Y-STR e Y-SNP para que estas possam ser utilizados para a análise quantitativa de incidências dos mesmos em casos de investigação forense; 2) a análise da estratificação da população masculina entre populações mundiais através da distribuição das frequências dos haplótipos Y-STR e Y-SNP; 3) a disponibilização de recursos para o estudo de Y-STRs e Y-SNPs.

Esta base de dados suporta os formatos haplotípicos mais comuns (Y Filer™ (Applied Biosystems) e PowerPlex® (Promega)), para os quais existem bases de dados de diferentes extensões. Estas são constituídas por informação populacional diretamente submetida por laboratórios individuais e que é verificada pelos membros da YHRD antes de ser disponibilizada para o público. Toda a informação populacional submetida para revistas forenses é previamente validada por depositários da YHRD e posteriormente publicados na própria base de dados.

A base de dados populacional da YHRD foi estruturada para que se possa estabelecer uma relação geográfica, filogenética e linguística entre os haplótipos pesquisados nesta mesma. A base de dados reconhece quatro metapopulações (grupos populacionais distribuídos definidos de acordo com um critério geográfico, étnico ou outro) que podem ser pesquisadas na plataforma: Nacional, Continental, Linguístico/Genético e Filogenético. Cada uma destas metapopulações contém várias subcategorias correspondentes.

No que toca à metapopulação nacional do YHRD, esta é composta por todos os indivíduos estudados em amostragens realizadas num determinado país, independentemente da origem dos mesmos (Willuweit e Roewer, 2015).

Metapopulações continentais são constituídas por todos os indivíduos cujas amostragens ocorreram num determinado continente de acordo com as definições geográficas das Nações Unidas. A base de dados YHRD permite então a pesquisa de haplótipos por continente, sendo estes: África, América Latina, América do Norte, Ártico, Ásia, Europa e Oceânia.

As metapopulações com base em etnicidade/linguística têm em conta a ascendência dos indivíduos em questão. Ascendência contém informação de várias categorias, incluindo cultural, geográfica, histórica e linguística. Estas categorias, apesar da sua caracterização ter algumas limitações, permite uma melhor descrição dos *clusters* genéticos observados e a sua relação com os padrões do cromossoma

Y encontrados. A hereditariedade linguística tem tendência a seguir padrões semelhantes à hereditariedade genética, pelo que esta é por vezes correlacionada com a transmissão de determinados padrões genéticos e polimorfismos ao longo de várias gerações. No entanto, esta categorização é posteriormente verificada com métodos estatísticos para confirmar as semelhanças e diferenças entre as várias amostras. Neste momento, a YHRD tem sete metapopulações diferentes pertencentes à metapopulação étnica/linguística: Aborígene Australiana, Africana, Afro-asiática, Ameríndia, Asiática Oriental, Esquimó-Aleúte e Euroasiática. Estas podem ainda apresentar subdivisões, como é o caso da metapopulação euroasiática que se divide em 6 subcategorias (Roewer *et al.*, 2005).

Todos os cromossomas Y que partilhem uma mutação são considerados relacionados por descendência até que uma nova mutação apareça. Haplótipos podem revelar-se extremamente semelhantes ou mesmo idênticos por descendência. Daí que se dê preferência a haplogrupos como critério para agrupar as amostras da base de dados por relações filogenéticas. Vários haplogrupos podem ser relacionados com eventos migratórios da pré-história humana (Underhill *et al.*, 2000).

A base de dados da YHRD é útil para a comparação das populações em estudo com a população portuguesa sendo que esta contém ferramentas para permitem estimar a distância genética entre estas. A distância genética é uma medida de divergência genética entre espécies ou populações de uma mesma espécie, e pode ser representada em termos de distância temporal do último ancestral comum ou em termos de grau de diferenciação. Populações com estruturas genéticas semelhantes entre si, apresentarão menores distâncias genéticas, o que significa que têm um ancestral comum recente e que têm proximidade genética.

A distância genética pode ser utilizada para a reconstrução de história populacional, bem como para a compreensão da origem da biodiversidade. Quando duas populações divergem e a sua *pool* genética se mantém separada por determinados fatores (normalmente geográficos), variações alélicas que apareçam posteriormente a esta separação, são específicas de cada população, aumentando a distância genética relativamente a outras. Sendo assim, a análise das frequências alélicas e haplotípicas e o posterior cálculo das distâncias genéticas entre populações podem ser utilizados para obter uma estimativa de quando esta separação ocorreu.

Existem várias medidas estatísticas para o cálculo de distância genética. O método mais frequentemente utilizado na estimativa a partir de informação de polimorfismos genéticos (como é o caso de SNPs e STRs) para o estudo de genética de populações é o índex de fixação (F_{ST}). Os valores de F_{ST} variam entre 0 e 1, sendo que quanto mais perto de 1 o valor, maior será a distância genética entre as populações, ou seja, maiores as diferenças entre os haplótipos das mesmas. F_{ST} está relacionado com a variância na frequência alélica entre as populações em estudo, sendo que se o seu valor é mais próximo de 0, isto significa que as frequências alélicas entre estas populações são semelhantes. Em populações entre as quais não existem grandes níveis de migração, os valores de F_{ST} tendem para 1, enquanto populações com maiores incidências de migração entre si, tendem a apresentar maiores semelhanças e daí menores valores de distância genética.

1.5. Objetivos do trabalho

Este estudo tem como objetivo a caracterização de marcadores genéticos do cromossoma Y na população de imigrantes africanos, utilizando o kit de amplificação PowerPlex® Y23 (Promega). Frequências haplotípicas bem como frequências alélicas serão estimadas com o objetivo de obter informação sobre a estrutura genética das populações imigrantes. A subsequente comparação entre as

populações imigrantes e a população de acolhimento permite avaliar o potencial impacto destes fluxos migratórios nas frequências alélicas encontradas na população, e consequente efeito nos exames genéticos realizados em contextos forenses.

Os resultados obtidos para os fins deste estudo poderão ser ainda inseridos na base de dados YHRD, para que estes possam vir a ser utilizados para futuros estudos populacionais ou forenses.

2. Materiais e Métodos

Foram analisadas um total de 400 amostras selecionadas aleatoriamente de entre imigrantes provenientes de países africanos, não relacionados entre si, residentes no sul de Portugal, e intervenientes em exames periciais que decorreram há mais de 2 anos no INMLCF-Delegação do Sul. Destas amostras, 201 pertenciam a indivíduos com origem cabo-verdiana, 85 com origem angolana, 61 de São Tomé e Príncipe, e 53 da Guiné-Bissau.

O DNA foi extraído a partir de manchas de sangue. A extração de DNA das amostras de sangue foi efetuada pelo método de Chelex 100® (Walsh, Metzger e Higuchi, 1991)..

Chelex 100 é um composto quelante, cujas propriedades permitem a purificação de outros compostos por troca iónica. Ao ligar-se a iões como o Mg^{2+}, responsável pela ativação de DNases e outras enzimas que contribuem para a degradação do DNA, este composto irá preservar as amostras para sua subsequente utilização em PCR. A resina é constituída por copolímeros estireno-divinilbenzeno com iões iminodiacetato, e mantêm componentes celulares sedimentados após a quebra das paredes celulares, enquanto o DNA e RNA mantêm-se em solução no sobrenadante da amostra.

O tempo de incubação a 56ºC após adição da solução de Chelex a 5% foi aumentado para 20 minutos na extração de amostras com tempo de armazenamento superior a 10 anos para garantir a presença de DNA, enquanto se manteve a 15 minutos para as restantes. No caso de ausência de manchas de sangue, foram extraídas amostras de DNA a partir de zaragatoas bucais utilizando o mesmo método de extração e seguindo as mesmas alterações de protocolo.

O DNA extraído das amostras foi quantificado utilizando o Quantifiler™ Trio DNA Quantification Kit (Applied Biosystems), de acordo com algumas modificações no procedimento sugerido pelo fabricante. Este kit utiliza como alvos *loci* multi-cópia para uma maior sensibilidade de deteção, sendo que estes consistem em múltiplas cópias dispersas por vários cromossomas autossómicos e pelo cromossoma Y, e estão descritos na Tabela 2.1 (Thermo Fisher Scientific, 2017).

Tabela 2.1-Alvos do kit Quantifiler™ Trio DNA Quantification Kit (Applied Biosystems). Fonte: Quantifiler™ HP and Trio DNA Quantification Kits, USER GUIDE (Thermo Fisher Scientific, 2017).

Alvo	Tamanho Amplicão	Ploidia	# de cópias	Marcador/Quencher
Alvo Humano, autossómico *small*	80 bases	Diploide	Multi-cópia	VIC™ dye with MGB quencher
Alvo Humano, autossómico *large*	214 bases	Diploide	Multi-cópia	ABY™ dye with QSY™ quencher
Alvo Humano Masculino	75 bases	Haploide	Multi-cópia	FAM™ dye with MGB quencher
Controlo Interno do PCR	130 bases	NA	Molde IPC sintético incluído na *primer mix*	JUN™ dye with QSY™ quencher

Neste estudo foram utilizadas metade das quantidades de reagentes sugeridas pelo fabricante, mantendo as proporções indicadas. Uma vez que a quantificação de DNA é um método muito sensível e com grande variabilidade, previamente a esta modificação foi efetuado um estudo com 100 amostras

selecionadas aleatoriamente. Foram efetuados os testes usando as condições do fabricante e reduzindo as quantidades do reagente a metade. Verificou-se através de um teste de ANOVA, que não existiam diferenças significativas entre as quantidades obtidas com metade dos reagentes e as condições do fabricante. O ensaio como o Quantifiler™ Trio DNA Quantification Kit foi realizado através da amplificação de fragmentos em tempo real no equipamento 7500 Real Time PCR System (Applied Biosystems). Este método calcula a concentração de DNA presente em cada uma das amostras a partir de uma curva de calibração realizada com 5 Standards com concentrações conhecidas, diluídas num fator de 10 (50 ng/µL; 5 ng/µL; 0,5 ng/µL; 0,05 ng/µL; 0,005 ng/µL). A cada ciclo de PCR, o número de cópias de DNA presente nos poços de reação aumenta exponencialmente a partir de uma concentração inicial de DNA na amostra, sendo esta relação calculada com base na curva de calibração usando regressão logarítmica implementada no software HID Real-Time PCR Analysis Software v1.2 (Thermo Fisher Scientific, 2017).

A informação obtida a partir da quantificação contribuiu diretamente para a decisão sobre qual o volume de amostra extraída a utilizar para a amplificação por PCR dos marcadores genéticos do cromossoma Y a serem estudados. Amostras com uma menor concentração de DNA foram amplificadas substituindo o volume de água utilizado por um volume maior de amostra até completar o volume total de reação para a amplificação de fragmentos de modo a obter a quantidade de DNA de cerca de 0,5 ng/µL por cada tubo de PCR.

Os marcadores genéticos do cromossoma Y incluídos no kit PowerPlex® Y23 System (Promega) foram co-amplificados num termociclador 9700 Perkin Elmer (Applied Biosystems), com um quarto das quantidades de reagentes sugeridas pelo fabricante, de acordo com as proporções indicadas, respeitando as proporções, num volume final de 6,5µL. A utilização destas quantidades de reagentes foi previamente testada no laboratório no qual este projeto foi realizado, de modo a garantir que não existem diferenças significativas nos resultados obtidos. O kit utilizado contém *primers* adequados para a amplificação simultânea de 23 marcadores genéticos do cromossoma Y através do método de PCR. Na tabela 2.2, estão referidas as condições do programa de amplificação. A análise de fragmentos foi efetuada por eletroforese capilar no sequenciador 3130 *xl* (Applied Biosystems) com formamida e o sizer interno indicado. A análise de produtos amplificados foi efetuada com o software GeneMapper® ID-X1.4 utilizando o marcador de peso molecular indicado para o kit de amplificação utilizado (Promega, 2017).

Tabela 2.2-Especificações do programa de amplificação em PCR utilizadas com o kit de amplificação PowerPlex® Y23 System (Promega).

PowerPlex® Y23 System		Temperatura (°C)	Tempo
Desnaturação inicial		96	2 minutos
30 ciclos	Desnaturação	94	10 segundos
	Annealing de primers	61	1 minuto
	Extensão	72	30 segundos
Extensão Final		60	20 minutos
Temperatura de conservação		4	∞

Em amostras com marcadores genéticos nos quais foram encontradas possíveis duplicações, deleções ou microvariantes raras, foi extraído o DNA e os marcadores amplificados novamente com um kit de amplificação de outro fabricante que apresenta *primers* diferentes. Não tendo sido possível efetuar a sequenciação das amostras em questão, a confirmação destes resultados foi realizada através de uma nova extração do DNA através do método Chelex 100® acima descrito, e foi feita uma nova co-

amplificação num termociclador 9700 Perkin Elmer (Applied Biosystems), utilizando o kit de amplificação Y Filer™ Plus (Applied Biosystems), de acordo com as instruções do fabricante (Thermo Fisher Scientific, 2014). Na tabela 2.3, estão referidas as condições do programa de amplificação. Este kit não contém *primers* para a amplificação do marcador genético DYS549, pelo que a confirmação de duplicações neste mesmo foram efetuadas por repetição da amplificação com o kit PowerPlex® Y23 System (Promega) após nova extração das amostras em questão com o método de Chelex 100®.

Tabela 2.3-Especificações do programa de amplificação em PCR utilizadas com o kit de amplificação Y Filer™ Plus (Applied Biosystems).

YFiler™ Plus		Temperatura (°C)	Tempo
Desnaturação inicial		95	1 minuto
30 ciclos	Desnaturação	94	4 segundos
	Annealing e Extensão	61,5	1 minuto
Extensão Final		60	22 minutos
Temperatura de conservação		4	∞

A diversidade de haplótipos e as distâncias genéticas entre populações selecionadas, bem como as frequências haplotípicas foram estimadas utilizando o software Arlequin 3.5.2.2. Este programa é um software gratuito, disponível para download para todos os sistemas operativos e que permite fazer análises de genética populacional.

As ferramentas do programa permitem a realização de vários testes, como é o caso de F-statistics (ou índex de fixação), desequilíbrio de *linkage* e *pairwise difference*, entre outros. Para os efeitos deste estudo, foram realizados alguns destes testes nas várias populações intervenientes, tendo sido calculadas as distâncias genéticas entre as mesmas. O programa estima as distâncias entre as populações através do número de alelos diferentes encontrados nos vários haplótipos.

Molecular Evolutionary Genetics Analysis (MEGA) é um software computacional gratuito utilizado para a análise estatística de evolução molecular e construção de árvores filogenéticas, incluindo vários métodos e ferramentas de filogenética. Este programa tem como *input* sequências (de DNA, RNA ou proteicas), no entanto pode calcular e trabalhar com matrizes de distância genética entre espécies ou, no caso deste estudo, populações. É possível então estimar uma árvore filogenética com as relações entre as populações em estudo de acordo com as distâncias genéticas estimadas pelo software Arlequin 3.5.2.2, criando uma árvore através do método *Neighbour-joining*, um método de *clustering* aglomerante (Saitou and Nei, 1987). Para os efeitos deste estudo, foi utilizada a versão 7.0 do programa, disponibilizada em Janeiro de 2016 (Kumar, Stecher and Tamura, 2016). A partir desta informação, podemos obter uma árvore filogenética que poderá elucidar as relações genéticas entre as populações envolvidas neste estudo e inferir uma evolução genética e geográfica da distribuição dos haplótipos encontrados.

A comparação entre as populações em estudo e a população de acolhimento foi efetuada recorrendo às ferramentas disponibilizadas pela base de dados internacional do cromossoma Y em https://yhrd.org. O método de Análise de Variância Molecular (AMOVA) utilizado considera a variância entre o número de repetições nos vários Y-STRs dentro e entre populações, tendo em conta a relação molecular entre os alelos e não apenas a sua frequência e possibilitando o cálculo de F_{ST} entre pares de populações (Excoffier, Smouse and Quattro, 1992). Baseando-se num algoritmo *Multidimensional Scaling* (MDS) não-métrico, a plataforma YHRD realiza um cálculo MDS para obter uma imagem representativa da matriz de distância entre as populações (Kruskal, 1964).

3. Resultados

Apesar de residirem em Portugal imigrantes originários de Moçambique (3,3% dos 18,8% totais de imigrantes de origem africana), o número de amostras disponíveis de indivíduos do sexo masculino com ascendência moçambicana não era suficiente para obter resultados com significado estatístico para este estudo. Foram então estudados apenas os marcadores de indivíduos com origem cabo-verdiana, angolana, guineense e são-tomense para o âmbito deste projeto. Em particular, foi extraído, quantificado e analisado o DNA de 201 amostras provenientes de indivíduos do sexo masculino de Cabo Verde, 61 de São Tomé, 53 de Guiné-Bissau e 85 de Angola com um total de 400 amostras. As percentagens de amostras em estudo não refletem perfeitamente as percentagens de imigrantes residentes encontradas em Portugal devido a limitações no número de amostras disponíveis, nomeadamente de indivíduos com origem guineense. Foram caracterizados os 23 marcadores do cromossoma Y cujos *primers* para amplificação por PCR estão incluídos no kit utilizado para este estudo, conforme referido na Introdução.

Foram calculadas as frequências dos vários alelos encontrados nos marcadores genéticos estudados do cromossoma Y. Na Figura 3.1, apresentam-se as frequências alélicas verificadas em todos os marcadores agrupados de acordo com o número de repetições dos motivos dos STR. As tabelas e respetivos gráficos das frequências alélicas separadas por marcador genético encontram-se em Anexo (Anexo 1 e 2).

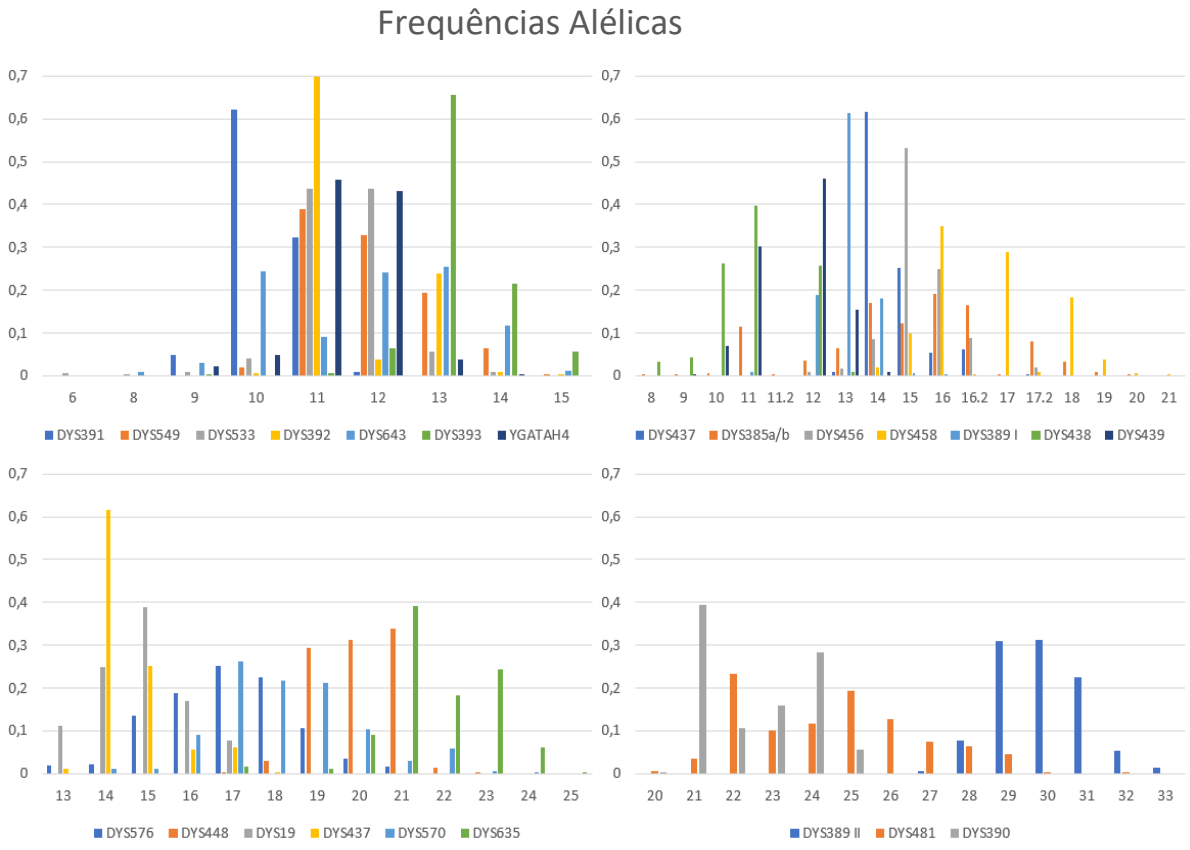

Figura 3.1-Gráficos de barras com as frequências alélicas para os marcadores genéticos do cromossoma Y estudados neste trabalho, por número de repetições dos motivos repetitivos de DNA em abcissas.

De entre as 400 amostras, foram encontrados 377 haplótipos diferentes, resultando numa diversidade haplotípica de 94,25%. Os haplótipos repetidos encontravam-se dentro das mesmas populações com exceção de um, cujo haplótipo é partilhado por um indivíduo imigrante de Cabo Verde e um indivíduo imigrante de Guiné-Bissau.

Um total de 4 indivíduos com origem de Cabo Verde partilham o mesmo haplótipo, apresentando uma frequência relativa de 0,0199 nesta população. Os restantes 12 haplótipos repetidos encontrados nas várias amostras cabo-verdianas apresentam frequências relativas de 0,00995 na mesma, e haplótipos únicos uma frequência relativa de 0,00498.

De entre os indivíduos imigrantes de Angola, apenas 2 apresentavam um haplótipo partilhado. Este haplótipo representa uma frequência relativa de 0,0235 dentro desta população, e os restantes haplótipos uma frequência de 0,0118 cada um.

Nas amostras estudadas de indivíduos provenientes de São Tomé e Príncipe, 1 haplótipo com três incidências tem uma frequência de 0,0492. Mais 4 haplótipos encontram-se repetidos nos vários indivíduos estudados, e apresentam uma frequência relativa de 0,0328. Os restantes haplótipos apresentam respetivamente frequências de 0,0164.

No caso das amostras estudadas de imigrantes de Guiné-Bissau, apenas 1 haplótipo se encontra repetido e tem uma frequência relativa de 0,0377 dentro da população de amostras. Os restantes haplótipos apresentam frequências de 0,0189, incluindo o haplótipo partilhado com um indivíduo de Cabo Verde, visto que este não é encontrado em mais nenhum indivíduo guineense. Devido à conservação do cromossoma Y de pai para filho, existirá uma relação de parentesco entre estes indivíduos, cujos parentes antepassados poderão ter migrado entre Guiné-Bissau e Cabo Verde, continuando a transmissão deste haplótipo numa nova população.

Entre todos os indivíduos estudados, cerca de 10% apresentavam variantes alélicas (diferentes formas dos alelos em estudo que não correspondem às variantes frequentemente encontradas nos alelos em estudo; e.g. duplicações de marcadores do cromossoma Y). Como se pode observar na Figura 3.2, 67% das variantes verificadas tratam-se de duplicações, 20% microvariantes raras e 13% alelos nulos.

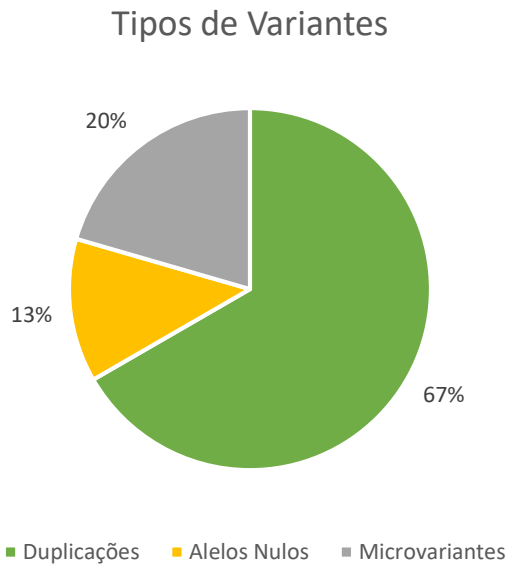

Figura 3.2-Percentagens dos tipos de variantes alélicas encontradas nas várias amostras estudadas.

As duplicações observadas verificaram-se apenas em três marcadores genéticos, sendo 85% das mesmas no marcador DYS448. Estas são na sua maioria duplicações 19,20, havendo também duas 20,21 e uma 19,21. Encontraram-se duplicações nos marcadores DYS389 II, das quais duas também apresentavam uma duplicação no marcador DYS439 (29,30 e 10,11 respetivamente), e ainda uma duplicação no marcador DYS549 (11,12). Com exceção da duplicação 19,21 no marcador DYS448, todas as duplicações observadas foram previamente descritas e estão presentes para consulta na base de dados *National Institute of Standards and Technology's STRBase database* (NIST) (Linstorm and Mallard). Mesmo após nova extração e amplificação com o kit de amplificação Y Filer™ Plus (Applied Biosystems), todas estas duplicações se mantiveram.

Os alelos nulos observados verificaram-se nos marcadores genéticos DYS448 e DYS392, sendo que estes marcadores não eram amplificados com qualquer um dos kits de amplificação utilizados. As microvariantes raras foram observadas apenas nos marcadores genéticos DYS458 e DYS385.

Dos países considerados, os indivíduos com origem em Guiné-Bissau apresentavam uma maior incidência de haplótipos com variantes alélicas (21%) em comparação com os restantes. Além disso, das variantes encontradas nestas amostras, 86% eram duplicações encontradas no marcador genético DYS448 (Figura 3.3). Destas, apenas uma não se trata de uma duplicação 19,20, a mais comum neste estudo, tratando-se de uma duplicação 20,21. As microvariantes encontram-se ambas no marcador DYS385.

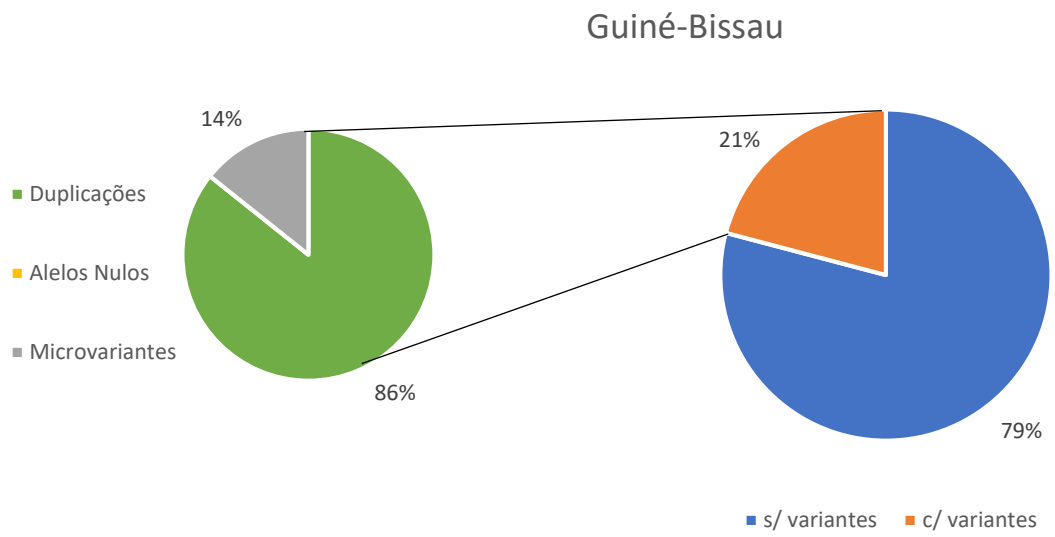

Figura 3.3-Percentagens relativas de variantes alélicas observadas nos haplótipos estudados de indivíduos imigrantes de Guiné-Bissau.

No caso de Cabo Verde, apesar de os indivíduos deste país terem sido estudados neste trabalho em maior número, estas amostras apresentam uma quantidade muito menor de variantes alélicas em relação a Guiné-Bissau, compreendendo cerca de 9% do total de amostras estudadas de indivíduos cabo-verdianos ou descendentes de cabo-verdianos. No entanto, a distribuição de variantes alélicas é a mais variada. De facto, das variantes encontradas, 65% tratam-se de duplicações, uma grande maioria em relação aos 20% e 15% de, respetivamente, microvariantes raras e alelos nulos (Figura 3.4).

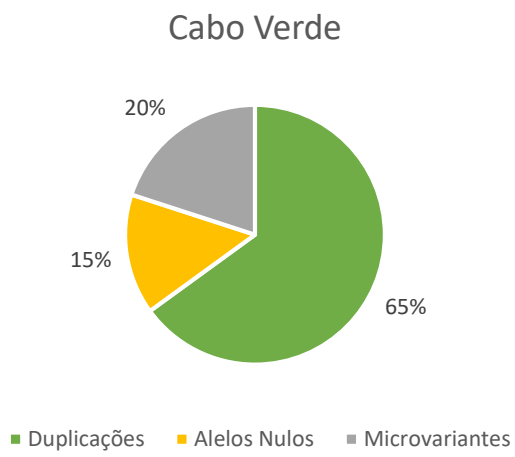

Figura 3.4-Percentagens observadas das variantes alélicas encontradas nos haplótipos de indivíduos Angolanos estudados.

A maioria das duplicações observadas nos haplótipos cabo-verdianos encontram-se no marcador genético DYS448, tal como nos restantes países, no entanto verificaram-se também duplicações nos marcadores DYS389 II e DYS439, como foi referido acima. Foi também nos haplótipos de Cabo Verde que se verificaram alelos nulos no marcador genético DYS392. Encontraram-se também microvariantes raras de vários tamanhos nos marcadores DYS458 e DYS385.

Nas amostras estudadas de indivíduos originários de São Tomé e Príncipe e Angola, a incidência de variantes alélicas era muito menos pronunciada (apenas 2 e 3 do total de amostras, respetivamente, que representam cerca de 3% do total estudado). Nas amostras de São Tomé e Príncipe, encontraram-se uma duplicação e uma microvariante rara, enquanto nas amostras de indivíduos angolanos foram encontrados dois alelos nulos e uma microvariante rara (Figura 3.5).

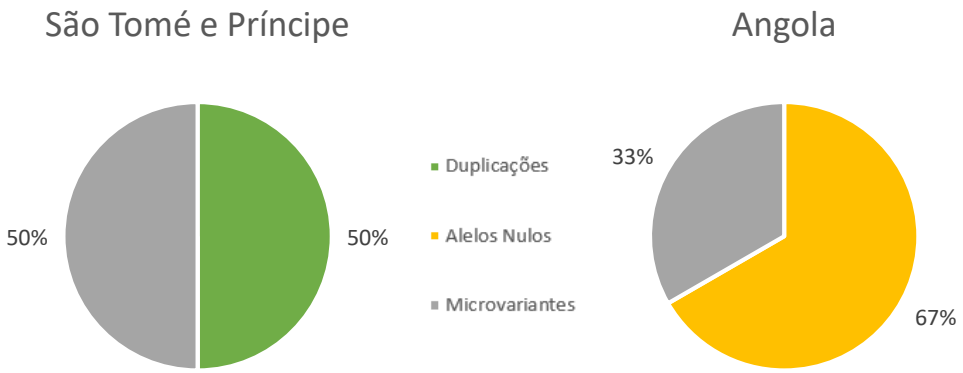

Figura 3.5-Percentagens observadas das variantes alélicas encontradas nos haplótipos de indivíduos originários de São Tomé e Príncipe e Angola estudados.

A duplicação encontrada entre as amostras de indivíduos de São Tomé e Príncipe encontra-se no marcador genético DYS549, bem como os alelos nulos encontrados nos haplótipos estudados de indivíduos Angolanos. Ambas as microvariantes tratam-se de um fragmento com 17.2 unidades de repetição no marcador genético DYS458.

Passando a uma análise comparativa entre as populações em estudo, a maior distância genética verifica-se entre as populações de Cabo Verde e Guiné-Bissau, com uma distância genética estimada de 0,03965. Segue-se a distância entre Cabo Verde e Angola, apresentando uma distância genética estimada

de 0,03307. A terceira maior distância genética verificou-se entre Guiné-Bissau e Angola, sendo esta 0,02154. Segue-se ainda a distância genética entre Guiné-Bissau e São Tomé e Príncipe, estimada em 0,02002, e as distâncias entre São Tomé e Cabo Verde e entre São Tomé e Angola, que se estimaram em, respetivamente, 0,01809 e 0,00210.

Utilizando o programa MEGA7, foi possível obter uma árvore filogenética a partir do método *Neighbour-Joining*, a qual pode ser observada na Figura 3.6.

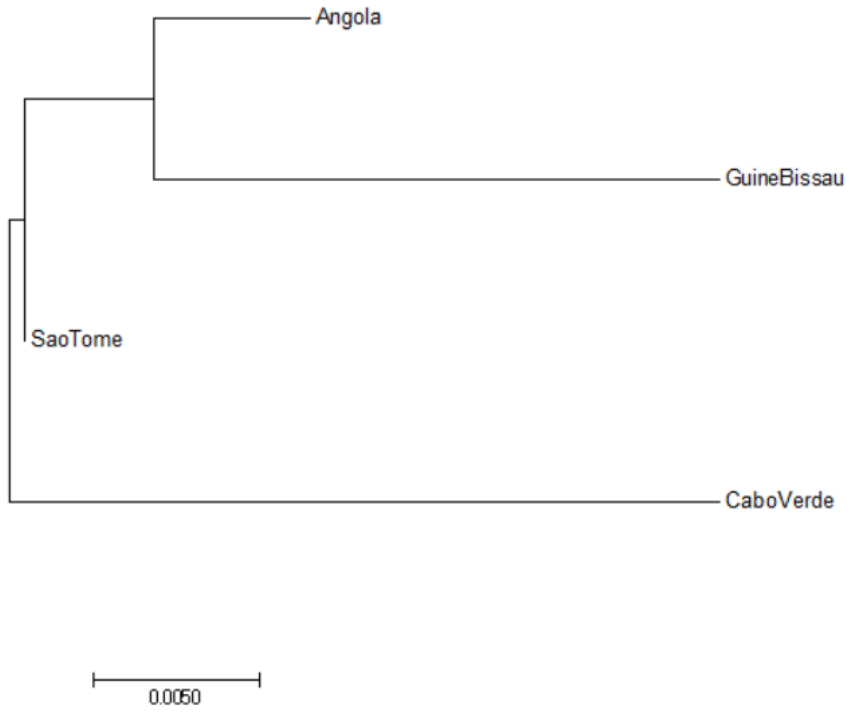

Figura 3.6-Árvore filogenética obtida utilizando o método Neighbour-Joining, a partir das distâncias genéticas estimadas pelo software Arlequin 3.5.2.2.

Analisando a árvore obtida, podemos verificar que, apesar de apresentarem distâncias genéticas estimadas elevadas, as populações de Angola e Guiné-Bissau são representadas no mesmo cluster. Cabo Verde e São Tomé e Príncipe encontram-se mais isoladas, mesmo havendo uma distância genética menor entre a população de São Tomé e Príncipe e todas as restantes.

Utilizando a base de dados YHRD e as suas ferramentas de análise e comparação entre populações através dos haplótipos observados, foi obtida uma imagem de MDS.

No entanto, as variantes alélicas observadas nas populações em estudo não foram reconhecidas pelas ferramentas de comparação entre populações do YHRD, pelo que estas informações não refletem a realidade de todas as amostras estudadas. Das 400 amostras originais, apenas 184 haplótipos de Cabo Verde, 83 de Angola, 60 de São Tomé e Príncipe e 41 de Guiné-Bissau foram incluídos na construção do gráfico MDS, e nenhum destes apresenta as variantes alélicas encontradas neste estudo, tendo sido excluídas as amostras com duplicações e alelos nulos.

Ainda assim, esta base de dados permite a comparação com haplótipos da população portuguesa previamente introduzidos na mesma. Apesar de não serem utilizados todos os haplótipos encontrados, a grande maioria dos mesmos ainda se encontra presente, pelo que a sua subsequente comparação com os haplótipos de indivíduos portugueses residentes no Sul e Centro de Portugal poderá contribuir com informações úteis para este estudo.

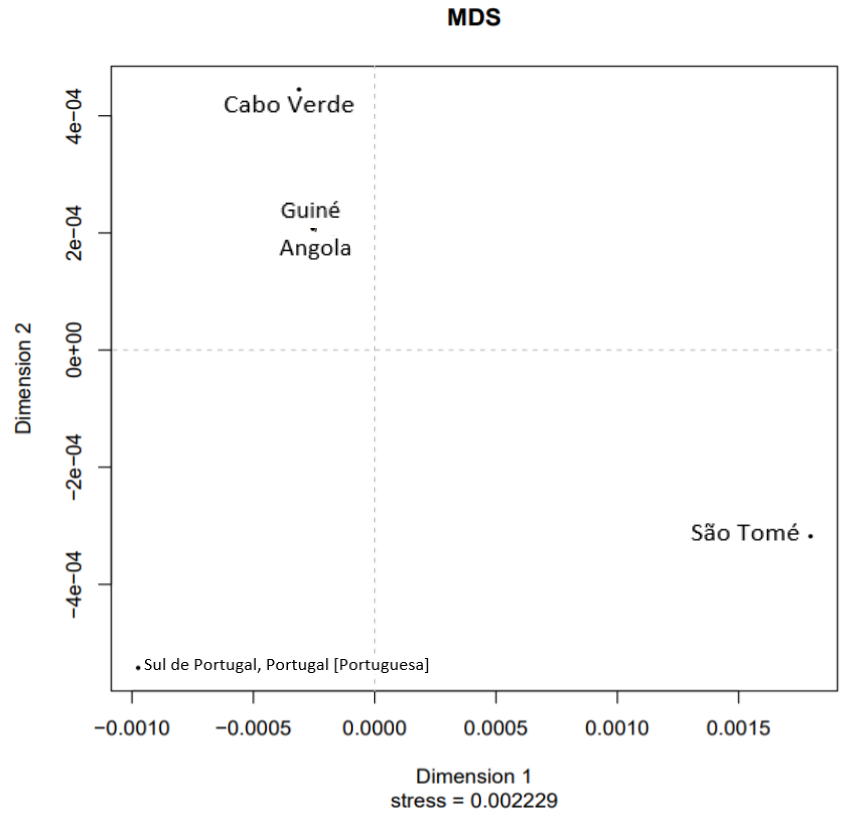

Figura 3.7-Gráfico MDS com a comparação entre os haplótipos encontrados nas amostras de Cabo Verde, Angola, São Tomé e Príncipe e Guiné-Bissau e a base de dados correspondente à população portuguesa do Sul de Portugal.

Como se pode observar na Figura 3.7, as populações de Guiné-Bissau e Angola são as mais próximas, tal como foi previamente descrito. No entanto, contrariamente ao que foi estimado pelo software Arlequin, a população de São Tomé e Príncipe encontra-se mais distante das restantes e não a população de Cabo Verde. Tal poderá ser devido aos haplótipos em falta, visto que o software Arlequin estima as distâncias genéticas de acordo com a presença de alelos semelhantes nos haplótipos em estudo. Uma vez que uma fração importante de haplótipos não é tida em consideração nos cálculos das distâncias em MDS, este método apresenta previsivelmente resultados que diferem da árvore filogenética gerada pelo MEGA7.

No entanto, tendo em conta a árvore filogenética representada na Figura 3.6, a presença das populações angolana e guineense no mesmo *cluster* bem como a separação das restantes, vêm de certa forma corroborar o que se encontra no gráfico MDS.

É ainda de notar que a população portuguesa se encontra fora de todos os quadrantes nos quais se verificam as populações africanas. Quer isto dizer que estas populações têm de facto diferenças significativas entre si e as suas frequências alélicas e estruturas genéticas deverão também ser diferentes. Dito isto, a inserção destes haplótipos na população poderá ter de facto um impacto significativo na estrutura genética da população residente no Sul e Centro de Portugal.

4. Discussão e Conclusões

Os resultados obtidos neste estudo vêm confirmar a existência de diferenças significativas, não só entre as populações em estudo, bem como entre estas e a população portuguesa de acolhimento.

As variantes alélicas observadas, quando inseridas na população de acolhimento, irão também afetar as frequências encontradas numa população a longo prazo, visto que estas são raras. Apesar de existirem ocorrências de variantes alélicas previamente descritas nos marcadores genéticos nos quais estas foram encontradas, a literatura descreve-as numericamente apenas na casa das unidades em cada estudo, sendo estas pouco frequentes. As amostras estudadas de indivíduos com origem em Angola e São Tomé e Príncipe verificaram-se nesta escala, no entanto, os haplótipos pertencentes a indivíduos de Cabo Verde e Guiné-Bissau, e em especial estes últimos, apresentavam uma percentagem de variantes alélicas muito maior do que seria de esperar de acordo com o número de amostras em estudo.

Existem registos de duplicações observadas nestes marcadores detetadas com o kit utilizado para este estudo. No entanto, é de notar que as ocorrências reportadas destas duplicações não são muito frequentes. Em particular, o marcador DYS448 apresenta duplicações com uma frequência muito mais elevada do que seria de esperar (presentes em cerca de 5,75% dos haplótipos em estudo). É possível que indivíduos originários de certos Países Africanos apresentem este polimorfismo em frequências mais elevadas, transmitindo esta duplicação ao longo de várias gerações. A sua introdução na população portuguesa pode vir a originar complicações na análise de amostras de origem desconhecida em exames de perícias forenses.

Dos marcadores genéticos nos quais foram encontradas variantes alélicas, é possível observar que os marcadores DYS389 II e DYS439 encontram-se bastante próximos no cromossoma Y (Figura 1.2). As duplicações observadas em ambos estes marcadores poderão dever-se a este facto, sendo possível que tenha ocorrido uma duplicação de parte deste segmento do cromossoma, levando à deteção simultânea de dois alelos nestes dois marcadores. No entanto, foi encontrado um haplótipo em que se verificou uma duplicação apenas no marcador DYS389 II. A duplicação deste segmento pode ainda assim ter ocorrido da mesma forma, no entanto o tamanho do segundo fragmento do marcador DYS439 pode ser igual ao primeiro. Neste caso, o pico verificado corresponde a duas cópias e não apenas a uma, tal como se verifica em haplótipos cujo marcador genético multi-cópia DYS385a/b apresenta apenas um pico no eletroferograma.

A verificação de múltiplos fragmentos num único marcador que não é considerado multi-cópia pode levar a complicações na análise de amostras em estudo em exames de perícias forenses, podendo resultar na identificação prematura de uma mistura de DNA masculino.

Devido à conservação do cromossoma Y ao longo de várias gerações de uma linha paterna com exceção de mutações pontuais, a possibilidade de perfis genéticos de cromossoma Y serem relativamente semelhantes entre dois indivíduos é muito mais elevada do que na análise de perfis autossómicos. Em situações nas quais as amostras de DNA se encontram demasiado degradadas para a obtenção de um perfil genético com marcadores autossómicos, a utilização de análise de marcadores do cromossoma Y é uma alternativa que pode reduzir a lista de suspeitos, logo é importante o conhecimento da existência destas variantes alélicas de modo a eliminar a possibilidade de haver uma mistura de DNA de forma mais rápida e eficaz.

Além disso, a caracterização e consideração destas duplicações cromossómicas pode servir para uma obtenção de perfis mais individualizantes, visto que estas variantes alélicas são relativamente raras na população em geral. Assim sendo, a presença das duplicações observadas neste estudo pode

efetivamente apontar para indivíduos que apresentam estas variantes e aumentar a probabilidade de coincidência com o perfil encontrado.

Para perceber o possível impacto das diferenças na estrutura genética das populações envolvidas em fluxos migratórios, é necessário compreender estas mesmas diferenças. As menores distâncias genéticas estimadas verificaram-se na sua totalidade entre a população de São Tomé e Príncipe e todas as restantes populações. É possível que estas ilhas tenham experienciado fluxos migratórios de várias áreas geográficas, resultando numa menor distância genética entre as populações envolvidas. Cabo Verde, apesar de ser também um grupo de ilhas, apresenta as maiores distâncias genéticas estimadas entre as restantes populações com exceção de São Tomé e Príncipe.

Figura 4.1-Imagem do mapa das fronteiras políticas dos vários países africanos, destacando os países dos quais as amostras estudadas têm origem. Fonte: https://pt.wikipedia.org/wiki/Ficheiro:Mapa_pol%C3%ADtico_da_%C3%81frica.svg. Acedido: Julho de 2019

A distância genética estimada entre as populações de Angola e Cabo Verde foi uma das maiores encontradas. É de notar que a distância geográfica entre estas populações também é a maior entre todas as estudadas (Figura 4.1). Esta distância poderá ter influenciado fluxos migratórios passados, havendo zonas geográficas mais próximas e mais atrativas para as populações migratórias e criado uma maior distância genética entre as populações de Angola e Cabo Verde. Da mesma forma, a distância genética estimada entre as populações de Angola e Guiné-Bissau também se verificou como uma das mais elevadas.

Como foi verificado pela árvore filogenética (Figura 3.6), a população de Cabo Verde mostra-se a mais isolada de todas as quatro em estudo. O facto de Cabo Verde e São Tomé e Príncipe se tratarem de grupos de ilhas é compatível com o que se observa nesta árvore filogenética. Na colonização de

pequenas ilhas, observa-se o efeito fundador, no qual a variabilidade genética da população colonizadora é apenas uma fração da variabilidade genética observada na população de origem. Em pequenas populações fundadoras, alelos que seriam raros na população de origem podem permanecer e proliferar na população colonizadora, gerando uma maior frequência destes mesmos ao longo das gerações (Cavalli-Sforza, Menozzi e Piazza, 1994). Tendo em conta este facto e ainda a colonização de Cabo Verde no século XV, isto poderá justificar as maiores distâncias genéticas com as restantes populações africanas. Fluxos migratórios relativamente constantes entre as populações continentais terão levado a alterações no seu *pool* genético, em contraste com a manutenção do *pool* genético das populações das ilhas. A maior incidência de duplicações no marcador genético DYS448 em indivíduos de Cabo Verde poderá ser um resultado do efeito fundador aquando da colonização destas ilhas.

Isto vem também corroborar a presença das populações angolana e guineense no mesmo *cluster*. Possíveis fluxos migratórios entre estas populações de África continental terão resultado em distâncias genéticas menores e maiores semelhanças nas frequências alélicas encontradas. No entanto, é de notar que ainda assim, a distância geográfica entre Angola e Guiné-Bissau é significativa, pelo que a migração poderá não ser muito elevada.

Um metaestudo populacional (Cavalli-Sforza, Menozzi e Piazza, 1994) no qual foram reunidos resultados de 49 populações para as quais se dispunha de mais informação genética, conduziu à construção duma árvore genética onde se puderam observar dois grandes *clusters*: as populações subsarianas e as populações de nordeste (Figura 4.2). Este último *cluster* subdivide-se ainda em dois *subclusters* (norte e este). As distâncias genéticas aqui representadas, apesar de semelhantes às estimadas neste projeto, não são comparáveis a estas últimas. Tal deve-se ao facto de este metaestudo utilizar informação genética de vários genes das populações, e não apenas marcadores do cromossoma Y, como é o caso dos resultados obtidos para os efeitos deste projeto.

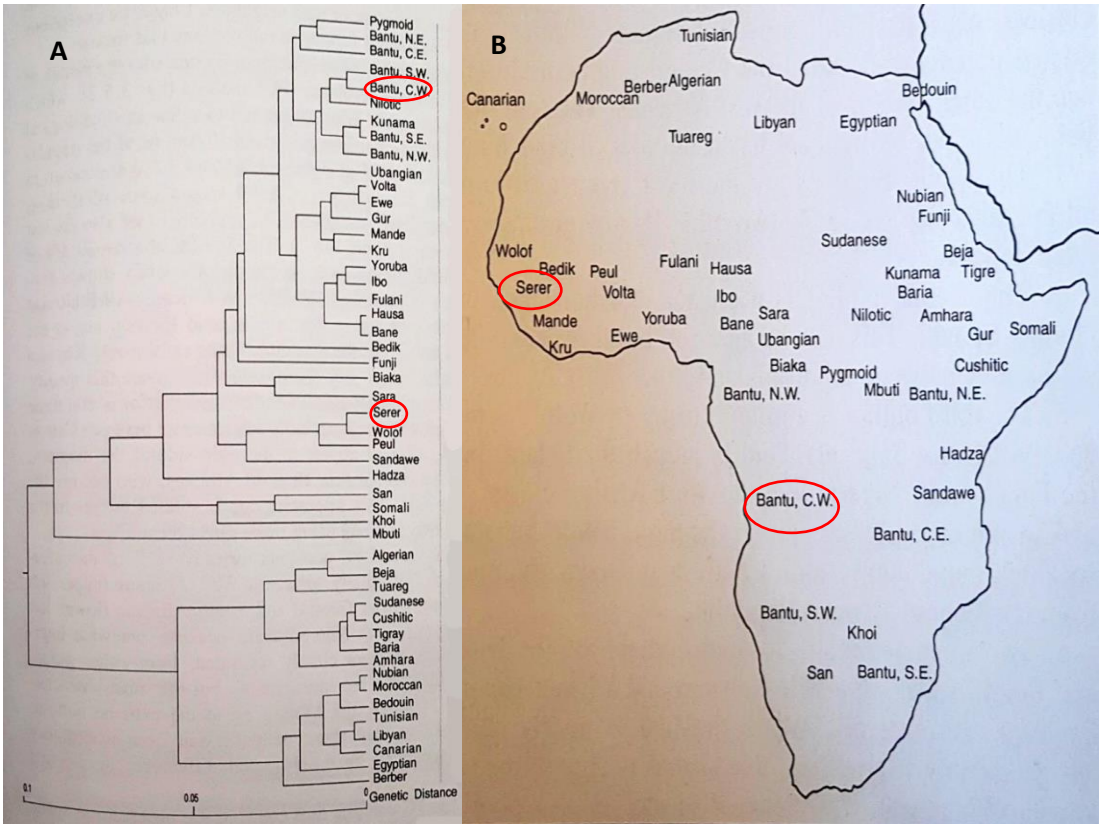

Figura 4.2-Imagens com populações africanas identificadas pelos seus países, grupos linguísticos ou outros. A: Imagem de uma árvore genética com 49 populações africanas. B: Imagem com as localizações geográficas de algumas das populações encontradas em A. Fonte: The History and Geography of Human Genes: Abridged Paperback Edition.

Ainda assim, é de notar que, no estudo metapopulacional, as populações genéticas descritas como correspondentes a Angola e Guiné, de acordo com as suas posições geográficas (Bantu C.W. e Serer), encontram-se ambas no *cluster* subsariano. Este facto vem corroborar a informação obtida na árvore filogenética na Figura 3.6, confirmando uma maior semelhança entre as populações de África Continental.

Comparando as distribuições das frequências alélicas dos 23 marcadores genéticos em estudo (Anexo 1), estas diferem das frequências presentes na base de dados YHRD para cada um dos marcadores. As frequências aí apresentadas correspondem a dados obtidos a nível mundial, enquanto as amostras analisadas neste estudo correspondem a uma área geográfica e origem étnica específicas. É de esperar que as frequências alélicas de determinadas populações difiram entre si, devido aos diferentes haplogrupos presentes nestas populações. Será expectável que os fluxos migratórios tenham impacto nas frequências alélicas das populações envolvidas (Cavalli-Sforza, Menozzi e Piazza, 1994).

Ao considerar que fluxos migratórios afetam as frequências alélicas e a estrutura genética de populações de acolhimento, este facto irá ser refletido na distância genética entre as populações envolvidas. Qualquer população de acolhimento que verifique grandes influxos migratórios sofrerá alterações consideráveis na sua estrutura e *pool* genético. Sendo o Centro e Sul de Portugal zonas atrativas para imigrantes, será de esperar que a inserção de populações migrantes irá afetar a estrutura genética da população portuguesa. A inserção de alelos anteriormente considerados raros ou mesmo inexistentes irá alterar as frequências relativas dos alelos presentes na população, o que irá afetar investigações no âmbito da genética forense.

Como referido anteriormente, o cálculo da probabilidade de perfis genéticos obtidos estarem relacionados é dependente das frequências populacionais dos alelos em estudo. A introdução de populações africanas na população portuguesa levanta a questão de quais frequências alélicas deverão ser utilizadas para este cálculo: africana ou portuguesa. Além disso, considerando o aumento da imigração em Portugal, alelos comuns nas populações imigrantes serão encontrados mais frequentemente em futuras gerações. Os cruzamentos entre homens de origem africana e mulheres portuguesas levarão a que as frequências alélicas referentes aos marcadores genéticos do cromossoma Y difiram das utilizadas atualmente, pelo que as probabilidades calculadas em exames de perícias forenses serão afetadas.

Os resultados aqui obtidos poderão servir como base para futuros estudos com o objetivo de caracterizar a população portuguesa em constante evolução, bem como avaliar mais concretamente o impacto da imigração na sua estrutura a partir das frequências alélicas aqui descritas. Os dados haplotípicos, caso sejam publicados na base YHRD, poderão ainda ser utilizados para outros estudos populacionais internacionais, oferecendo informação mais correta no que respeita à atual estrutura da população portuguesa.

5. Bibliografia

Balanovsky, O. (2017) 'Toward a consensus on SNP and STR mutation rates on the human Y-chromosome', *Human Genetics*. Springer Berlin Heidelberg, 136(5), pp. 575–590. doi: 10.1007/s00439-017-1805-8.

Brinkmann, B. *et al.* (1998) 'Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat', *American Journal of Human Genetics*, 68, pp. 1408–1415. doi: 10.1086/301869.

Cavalli-Sforza, L. L., Menozzi, P. and Piazza, A. (1994) *The History and Geography of Human Genes: Abridged Paperback Edition*. Abridged P. Princeton, New Jersey: Princeton University Press.

Excoffier, L., Smouse, P. E. and Quattro, J. M. (1992) 'Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application', *Genetics*, 131(2), pp. 479–491.

Gill, P., Jeffreys, A. J. and Werrett, D. J. (1985) 'Forensic application of DNA "fingerprints"', *Nature*, 318, pp. 577–579.

Glover, K. A. *et al.* (2010) 'A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment', *BMC Genetics*, 11(2). doi: 10.1186/1471-2156-11-2.

Hedrick, P. (2011) *Genetics of Populations*. 4th Edition. Tempe, Arizona: Jones & Bartlett.

Kruskal, J. B. (1964) 'Nonmetric multidimensional scaling: a numerical method', *Psychometrika*, 29(2), pp. 115–129.

Kumar, S., Stecher, G. and Tamura, K. (2016) 'MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets', *Molecular Biology and Evolution*, 33(7), pp. 1870–1874. doi: 10.1093/molbev/msw054.

Linstorm, P. J. and Mallard, W. G. (eds) (no date) *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. Gaithersburg MD: National Institute of Standards and Technology. doi: https://doi.org/10.18434/T4D303.

Meyer, E. *et al.* (1995) 'Microsatellite polymorphisms reveal phylogenetic relationships in primates', *Journal of Molecular Evolution*, 41(1), pp. 10–14. doi: 10.1007/BF00174036.

PORDATA (2018) *Base de Dados de Portugal Contemporâneo*. Available at: https://www.pordata.pt/Portugal (Accessed: 23 September 2019).

Promega (2017) *PowerPlex ® Y23 System for Use on the Applied Biosystems ® Genetic Analyzers*. Available at: https://worldwide.promega.com/-/media/files/resources/protocols/technical-manuals/101/powerplex-y23-system-protocol.pdf?la=en.

Roewer, L. *et al.* (2005) 'Signature of recent historical events in the European Y-chromosomal STR haplotype distribution', *Human Genetics*, 116, pp. 279–291. doi: 10.1007/s00439-004-1201-z.

Roewer, L. (2009) 'Y chromosome STR typing in crime casework', *Forensic Science, Medicine, and Pathology*, 5(2), pp. 77–84. doi: 10.1007/s12024-009-9089-5.

Saitou, N. and Nei, M. (1987) 'The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees', *Molecular Biology and Evolution*, 4(4), pp. 406–425.

Tautz, D. and Renz, M. (1984) 'Simple sequences are ubiquitous repetitive components of eukaryotic genomes', *Nucleic Acids Research*, 12(10), pp. 4127–4138. doi: 10.1093/nar/12.10.4127.

Thermo Fisher Scientific (2014) *Yfiler ™ Plus PCR Amplification Kit USER GUIDE*. Available at: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4485610_YfilerPlus_UG.pdf.

Thermo Fisher Scientific (2017) *Quantifiler ™ HP and Trio DNA Quantification Kits USER GUIDE*.

G. Available at: http://tools.thermofisher.com/content/sfs/manuals/4485354.pdf.

Underhill, P. A. *et al.* (2000) 'Y chromosome sequence variation and the history of human populations', *Nature Genetics*, 26, pp. 358–361.

Walsh, P. S., Metzger, D. A. and Higuchi, R. (1991) 'Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material', *BioTechniques*, 10(4), pp. 506–513. doi: 10.2144/000114018.

Weber, J. L. and May, P. E. (1989) 'Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction.', *American journal of human genetics*, 44(3), pp. 388–96. doi: 10.1146/annurev.anchem.1.031207.112938.

Willuweit, S. and Roewer, L. (2015) 'The new Y Chromosome Haplotype Reference Database', *Forensic Science International: Genetics*. Elsevier Ireland Ltd, (15), pp. 43–48. doi: 10.1016/j.fsigen.2014.11.024.